Discriminative training of CRF models with probably submodular constraints

Wojciech Zaremba
Dept. of Computer Science
Courant Institute
New York University
woj.zaremba@gmail.com

Matthew B. Blaschko
Center for Processing Speech and Images
Dept. of Electrical Engineering
KU Leuven
matthew.blaschko@kuleuven.be

Abstract

Problems of segmentation, denoising, registration and 3d reconstruction are often addressed with the graph cut algorithm. However, solving an unconstrained graph cut problem is NP-hard. For tractable optimization, pairwise potentials have to fulfill the submodularity inequality. In our learning paradigm, pairwise potentials are created as the dot product of a learned vector w with positive feature vectors. In order to constrain such a model to remain tractable, previous approaches have enforced the weight vector to be positive for pairwise potentials in which the labels differ, and set pairwise potentials to zero in the case that the label remains the same. Such constraints are sufficient to guarantee that the resulting pairwise potentials satisfy the submodularity inequality. However, we show that such an approach unnecessarily restricts the capacity of the learned models.

Instead, we approach the problem of learning with submodularity constraints from a probabilistic setting. Prediction errors may be the result of: learning error, model error, or inference error. Guaranteeing submodularity for all possible inputs, no matter how improbable, reduces inference error to effectively zero, but increases model error. In contrast, we relax the requirement of guaranteed submodularity to solutions that are submodular with high probability. We show that the conceptually simple strategy of enforcing submodularity on the training examples guarantees with low sample complexity that test images will also yield submodular pairwise potentials. Results are presented showing substantial improvement from the resulting increased model capacity.

1. Introduction

Multiple problems emerging in computer vision, such as segmentation, denoising, registration and 3d reconstruction, are addressed with Structured Output Support Vector Machines (SSVM) applied to conditional random field (CRF) models. The arrising problem of energy minimization in CRFs can be solved by a variety of methods, including loopy belief propagation, alpha-expansion, alpha-beta swap and many others. A majority of energy minimalization algorithms require pairwise potentials to fulfill (pairwise) submodular constraints or metric constraints. This requirement places a strong limitation on the family of models that can be employed.

It is well known from statistical learning theory that the prediction error of a discriminant function can be decomposed into the error resulting from the learning procedure, and the error resulting from the model class [4]. In a structured output setting, such as in learning the parameters of a CRF model, a method may also have error resulting from suboptimal inference. In this work we explore the tradeoffs resulting from this third source of error, showing that (*a*) increasing model capacity by allowing some test-time potentials to be potentially non-submodular generally improves accuracies over guaranteeing submodularity for all possible inputs and (*b*) we can bound the probability of a non-submodular constraint occurring at test time with low sample complexity. This latter result indicates that relaxing submodularity constraints to guarantee "only" probably submodular potentials is a safe and principled strategy for increasing model capacity and increasing the resulting system accuracy.

In this work, we make several fundamental contributions to discriminative learning of CRF models: (*a*) a formulation for learning probably submodular constraints, (*b*) an algorithm for efficiently generating the most violated submodularity constraint, (*c*) the concept of a tradeoff between model error and inference error in CRF training, and (*d*) empirical results showing substantial improvement on a benchmark dataset.

1.1. Related work

Random field models in image segmentation initially employed data independent pairwise terms encoding a relatively simple prior that adjacent pixels were likely to have

the same label [10, 6]. The first data dependent pairwise terms proposed in the literature were simple contrast dependent terms with fixed positive weighting, resulting in a guarantee of submodularity [5]. In the first applications of structured output support vector machines to the discriminative learning of pairwise terms, only associative potentials were employed, enforced by a single positive constraint [2]. A later work employed only two positively constrained learned weights: one for a Potts-like term, and one for a contrast dependent term [15]. This simple positivity constraint is sufficient to guarantee submodularity for all possible inputs, but does not give the learning algorithm much capacity to optimize the pairwise terms. In contrast, we consider here the optimization of hundreds or thousands of pairwise parameters, providing a rich model space for learning informative pairwise potentials. An alternative approach is to consider only tree structured models [14], but this again restricts the model space and disallows potentially helpful model interactions.

In relaxing the constraint set to include models that do not guarantee submodularity for all possible inputs, we develop bounds on the probability of a test time input resulting in a non-submodular potential. This problem reduces to the problem of estimating the sample complexity of learning a convex cone by an intersection of half spaces. This problem is a central open question in computational learning theory, and existing results require strong assumptions on the generating distribution such as zero-mean log concavity [12], or the existence of a margin between positive and negative samples [3]. We take the comparatively conservative approach of upper-bounding the probability of lying outside a convex cone by the probability of lying outside a convex hull, and we present results that have a term depending on a moment functional of the underlying distribution [7]. We note that improvements on bounds available in this area of research will directly be applicable to the learning setting considered here.

The approach of bounding the error of an algorithm is closely related to the notion of probably approximately correct (PAC) learning [17]. In analogy to PAC learning, probabilistic bounds have been considered before in the development of inference algorithms for computer vision problems, such as in the development of thresholds for an object detection cascade architecture [9].

2. Discriminative Learning of Segmentation Models

A structured output support vector machine (SSVM) is an extension of the well-known support vector machine (SVM) [8] classifier, which enables the prediction of complex and interdependent outputs. Formally, let $x \in \mathcal{X}$ denote an input to be assigned an output $y \in \mathcal{Y}$, which represents in our case image segmentation. We assume a training set of labeled examples $\mathcal{S} \equiv \{(x_i, y_i)\}_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$. We represent the joint feature vector of an input x_i and output variable y_i by $\phi(x_i, y_i)$. Given a training dataset $\mathcal{S}$, the parameters w of the SSVM are learned by solving the following optimization problem [16]:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{1}$$

$$\text{s.t.} \langle w, \phi(x_i, y_i) - \phi(x_i, \tilde{y})\rangle - \Delta(y_i, \tilde{y}) \geq -\xi_i, \ \forall i \tilde{y} \tag{2}$$

$$\xi_i \geq 0, \ \forall i \tag{3}$$

$$w \in \mathcal{C} \tag{4}$$

The number of constraints (2) is large. They consist of all possible assignments of $\hat{y}$ for every sample (precisely $|\mathcal{Y}| \times n$ constraints). However, a cutting plane approach enables tractable optimization of this objective in a wide variety of settings [11]. We have chosen margin rescaling instead of slack rescaling here for simplicity of derivation.

The additional constraints in (4) ensure that the model results in tractable inference problems and are application specific. We will consider various forms of $\mathcal{C}$ in the sequel resulting in different model classes and inference guarantees. In the case of segmentation, these constraints are designed to ensure submodularity of the pairwise potentials, which is critical for energy minimalization using the graph cut algorithm.

It is straightforward to map the parameters of a log-linear graphical model to a joint feature map ϕ [16]. In the case of a conditional random field, we have [15]

$$-\langle w, \phi(x, y)\rangle = \sum_{k \in V} U(x^k, y^k, w) \ + \sum_{(k,l) \in E} P(x^k, y^k, x^l, y^l, w) \tag{5}$$

where $k \in V$ sums over pixels (vertices), and $(k, l) \in E$ sums over neighboring pairs of pixels (edges). As the terms U and P are linear, we have that the energy can be written

$$\langle w, \phi(x, y)\rangle = \left\langle \begin{pmatrix} w_u \\ w_p \end{pmatrix}, \begin{pmatrix} \sum_{k \in V} \phi_V(x^k, y^k) \\ \sum_{(k,l) \in E} \phi_E(x^k, y^k, x^l, y^l) \end{pmatrix} \right\rangle \tag{6}$$

It is well known that arbitrary, unconstrained unary potentials lead to efficient inference in graphical models, and that it is only the pairwise potentials that effect the tractability of the solution [6].

We decompose our pairwise feature function ϕ_E as a Kronecker product [13] over features of the pairs of pixels ϕ_x and of the labels ϕ_y

$$\phi_E(x^k, y^k, x^l, y^l) = \phi_y(y^k) \otimes \phi_y(y^l) \otimes \phi_x(x^k, x^l). \tag{7}$$

Here, we assume that $\phi_y : \mathcal{L} \mapsto \{0, 1\}^{|\mathcal{L}|}$ is an indicator function specifying the desired label in a label set $\mathcal{L}$, while

ϕ_x can be any positive[1] vector valued function measuring statistics of the difference between two (super-)pixels. Specializing this notion to the binary setting,[2] we note that this is equivalent to learning four separate weight vectors, which we will denote w_{00}, w_{01}, w_{10}, and w_{11}, resulting in a matrix of pairwise potentials

$$-\begin{pmatrix} \langle w_{00}, \phi_x(x^k, x^l)\rangle & \langle w_{01}, \phi_x(x^k, x^l)\rangle \\ \langle w_{10}, \phi_x(x^k, x^l)\rangle & \langle w_{11}, \phi_x(x^k, x^l)\rangle \end{pmatrix} \quad (8)$$

This matrix of potentials leads to a submodular inference problem on a given image iff

$$\langle w_{11}, \phi_x(x^k, x^l)\rangle + \langle w_{00}, \phi_x(x^k, x^l)\rangle - \langle w_{01}, \phi_x(x^k, x^l)\rangle - \langle w_{10}, \phi_x(x^k, x^l)\rangle \geq 0 \quad \forall k, l. \quad (9)$$

With this notation, we may now consider constraint sets, $\mathcal{C}$, that ensure that the inequalities in Equation (9) are satisfied.

2.1. Definitely Submodular Constraints

Szummer et al. proposed a simple set of positivity constraints [15] to ensure that the constraints in Equation (9) are guaranteed to be satisfied over the entire image of ϕ_x, that is *by all possible inputs, x, regardless of how improbable they may be*. In our notation, this constraint set is equivalent to the following

$$\mathcal{C}_1 \equiv \{w \mid w_{00} = w_{11} = \mathbf{0} \wedge w_{01} \preccurlyeq \mathbf{0} \wedge w_{10} \preccurlyeq \mathbf{0}\} \quad (10)$$

where $\preccurlyeq$ denotes element-wise inequality.

It is immediately clear on inspection of Equations (9) and (10) that this constraint set is sufficient, but not necessary. We therefore consider a second constraint set that is necessary and sufficient to guarantee submodularity for the entire image of ϕ_x.

$$\mathcal{C}_2 \equiv \{w \mid w_{00} \succcurlyeq \mathbf{0} \wedge w_{11} \succcurlyeq \mathbf{0} \wedge w_{01} \preccurlyeq \mathbf{0} \wedge w_{10} \preccurlyeq \mathbf{0}\}. \quad (11)$$

The set of models defined by $\mathcal{C}_2$ is strictly larger than the set defined by $\mathcal{C}_1$.

2.2. Probably Submodular Constraints

We now make a probabilistic argument, which we make precise in Section 3, that we may further relax $\mathcal{C}_2$ to enforce linear constraints on w of the form in Equation (9) only for the values of $\phi_x(x^k, x^l)$ occuring in the training data. We will refer to this constraint set as

$$\mathcal{C}_4 \equiv \{w \mid \langle w_{00}, \phi_x(x^k, x^l)\rangle + \langle w_{11}, \phi_x(x^k, x^l)\rangle - \langle w_{01}, \phi_x(x^k, x^l)\rangle - \langle w_{10}, \phi_x(x^k, x^l)\rangle \geq 0 \quad \forall k, l, x \in \mathcal{S}\} \quad (12)$$

The key insight that allows us to make this relaxation is that if a function is submodular on the training data, with high probability it will be submodular on the test data (see Section 3). Furthermore, the constraints are linear in w and our optimization remains a quadratic programming problem, albeit with a large constraint set.

As a final constraint set, we slightly restrict $\mathcal{C}_4$ to ensure that pairwise potentials of the same label are negative (i.e. favored by the inference procedure), while pairwise potentials of different labels are positive (i.e. discouraged by the inference procedure):

$$\mathcal{C}_3 \equiv \{w | \langle w_{00}, \phi_x(x^k, x^l)\rangle \geq 0 \wedge \langle w_{00}, \phi_x(x^k, x^l)\rangle \geq 0 \wedge \langle w_{01}, \phi_x(x^k, x^l)\rangle \leq 0 \wedge \langle w_{10}, \phi_x(x^k, x^l)\rangle \leq 0 \quad \forall k, l, x \in \mathcal{S}\}. \quad (13)$$

This specifies in a loose way prior knowledge about the role of pairwise constraints in image segmentation, while still giving sufficient model capacity to the learning algorithm.

We have that $\mathcal{C}_1 \subset \mathcal{C}_2 \subseteq \mathcal{C}_3 \subseteq \mathcal{C}_4$. Thus, we strictly increase the model capacity when we move from $\mathcal{C}_1$ to $\mathcal{C}_4$. $\mathcal{C}_4$ may in the limit reach $\mathcal{C}_2$, but this would require a very unnatural data set to impose such strong constraints. In all experiments, we observe that $\mathcal{C}_3$ and $\mathcal{C}_4$ are substantially larger than $\mathcal{C}_2$ and that the optimal weight vector achieved by the objective in Equation (1) optimized with constraints $\mathcal{C}_3$ lies outside $\mathcal{C}_2$. Similarly, we empirically observe that $\mathcal{C}_3$ and $\mathcal{C}_4$ are distinct and result in different optimal w (Section 5).

3. Sample Complexity of Probably Submodular Constraints

We consider that our training images x be drawn i.i.d. from some probability distribution $p(x)$ (this assumption is already implicit in the regularized risk minimization of the SSVM). We therefore consider the vector valued random variable $\phi_x(x_i^k, x_i^l)$ where x_i is drawn from $p(x)$ and k and l are sampled uniformly. Our precise task is to determine whether $\mathcal{C}_3$ and $\mathcal{C}_4$ determined by the training sample results in a high probability of the scalar random variable $\langle w_{00}, \phi_x(x_i^k, x_i^l)\rangle + \langle w_{11}, \phi_x(x_i^k, x_i^l)\rangle - \langle w_{01}, \phi_x(x_i^k, x_i^l)\rangle - \langle w_{10}, \phi_x(x_i^k, x_i^l)\rangle$ being non-negative, where w satisfies $\mathcal{C}_3$ or $\mathcal{C}_4$, respectively. We note that $\mathcal{C}_3$ and $\mathcal{C}_4$ are both convex cones as they are the intersection of half-spaces that intersect the origin. If we assume that ϕ_x is positive, this is equivalent to the open problem in statistical learning theory of the sample complexity of approximating a convex cone by the intersection of a finite set of half spaces. Although some precise results are known, they make assumptions on the generating distribution, such as zero-mean log-convexity [12] or the existence of a margin [3]. Here, we consider a conservative bound by noting that a convex cone enclosing the data is strictly larger than its convex hull and therefore the integral of the probability measure outside the

[1] We may relax the positivity requirement on ϕ_x in the probably submodular case, but it is required in the settings that guarantee submodularity for all possible inputs. One possible choice is an element-wise absolute difference of two feature vectors computed at sites x^k and x^l.

[2] All results derived here hold also for the multi-class setting with α-β swap optimization [6] by considering submodularity constraints to hold for all label pairs α and β.

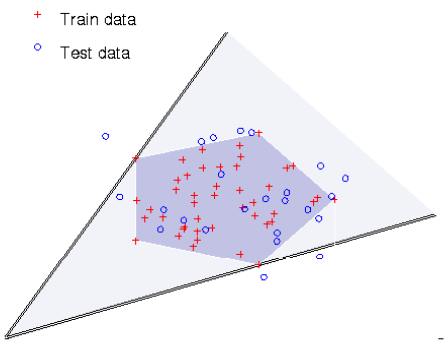


Figure 1: The probability of a test pairwise potential being submodular can be reduced to the question of the sample complexity of learning convex cones. To bound the probability of a random variable landing outside the convex cone defined by the training data, we use a bound on the probability of the random variable landing outside the convex hull. This is sufficient to bound the sample complexity of probably submodular constraints as defined in Section 2.2.

convex cone is strictly smaller than the integral of the probability measure outside the convex hull (Figure 1).

We additionally allow our sample complexity bound to contain terms dependent on a moment generating function of the underlying distribution on $\phi_x(x_i^k, x_i^l)$ rather than make restrictive assumptions on the distribution to eliminate this dependency.

Lemma 3.1 *Denote by V_n^d expected volume of the convex hull of n points chosen independently according to given distribution μ. One has*

$$V_{d+1+p} = \binom{d+1+p}{d+1} L(x^p + (1-x)^p) \qquad (14)$$

where L is a moment functional for distribution μ.

A proof can be found in [7]. For well-behaved distributions such as those found in natural image datasets, the underlying distribution has bounded higher moments and the latter term converges to zero quickly with p, indicating that V converges to one. This indicates that the constraints being satisfied on the training data will result in the constraints being satisfied on the test data with high probability.

4. Cutting Plane Generation of Most Violated Submodularity Constraint

In this section, we describe how to incorporate the large number of linear constraints that define $\mathcal{C}_3$ and $\mathcal{C}_4$ into the optimization algorithm. The number of constraints is proportional to the number of pairwise constraints in an image multiplied by the number of images. In a multi-class setting with submodularity constraints between all pairs of classes, the number of constraints grows quadratically with the number of classes. Analogous to Section 2, we will use the notation $w_{\alpha\alpha}$, $w_{\beta\beta}$, $w_{\alpha\beta}$, and $w_{\beta\alpha}$ to denote the four components of the vector for each pair of labels α and β.

The probably submodular formulation enforces the constraints :

$$\begin{aligned} &\langle w_{\alpha\alpha}, \phi_x(x^k, x^l)\rangle + \langle w_{\beta\beta}, \phi_x(x^k, x^l)\rangle \qquad (15)\\ &- \langle w_{\beta\alpha}, \phi_x(x^k, x^l)\rangle - \langle w_{\alpha\beta}, \phi_x(x^k, x^l)\rangle \geq 0 \quad \forall k, l. \end{aligned}$$

This set of constraints has to be satisfied for the every edge (x^k, x^l) in every training example. Moreover, it has to be satisfied for the every value of α, β. In theory, one may enumerate a matrix of all constraints on w, but it is prohibitive to do so. Instead, we will use a cutting plane approach and generate only the hard constraints on w that are violated at some point during the optimization. In practice, we observe that the number of such constraints is feasible. Nevertheless, we require an efficient method for generating the most violated constraint that does not require the explicit enumeration of all constraints.

Let us denote by P all pairwise potentials in the entire set of training data: $P = \{\phi_x(x^k, x^l)\}_{k,l \in E}$. The total number of constraints is $|P| \cdot \frac{|\mathcal{L}|(|\mathcal{L}|+1)}{2}$, and every entry is of the length of w. The matrix of constraints can be written as a Kronecker product of smaller matrices, which enables us to avoid storing a prohibitively large constriant matrix.

Let $B \in \{0,1\}^{\binom{|\mathcal{L}|}{2} \times |\mathcal{L}|^2}$ be a matrix with rows equal to

$$\begin{aligned} B_{i,:} =& (\phi_y(\alpha) \otimes \phi_y(\alpha))^T + (\phi_y(\beta) \otimes \phi_y(\beta))^T \qquad (16)\\ &- (\phi_y(\alpha) \otimes \phi_y(\beta))^T - (\phi_y(\beta) \otimes \phi_y(\alpha))^T \end{aligned}$$

for every assignment of α and β. The constraints (15) are of the form $(B \otimes P)w \geq \mathbf{0}$. Using Theorem 2.2 of [13] we may express these constraints by

$$(B \otimes P)w = \mathrm{vec}(P\tilde{w}B') \qquad (17)$$

where $\tilde{w}$ is a matrix of appropriate size such that $\mathrm{vec}\,\tilde{w} = w$. The computation on the right hand side of Equation (17) is substantially more memory and computation efficent than the left hand side. This enables the quick determination of the most violated hard constraint, which can then be added to the active set of constraints in a cutting plane optimization scheme.

5. Results

We have evaluated our method on the TU Darmstadt cows dataset. We have first over segmented images with SLIC superpixels [1]. Next, we have computed color histograms with 100 bins for every SLIC superpixel. We have used these features as the basis for our unary potentials,

	Active constraints	All constraints
Definitely submodular models		
$\mathcal{C}_0$	326	415
$\mathcal{C}_1$	301	426
$\mathcal{C}_2$	256	405
transductive $\mathcal{C}_4$	126	814
Probably submodular models		
$\mathcal{C}_3$	999	1609
$\mathcal{C}_4$	116	699

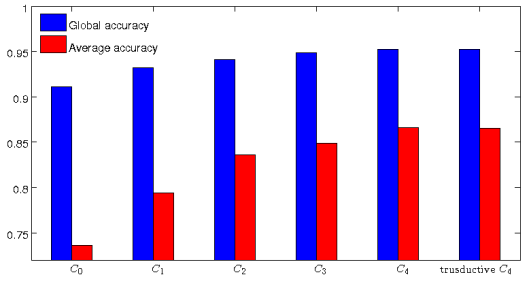
Table 1: The number of active constraints at the convergence point and the number of all constraints that were active at some point during optimization (cf. Section **??**).

and their absolute value of the difference in features as the basis for our pairwise potentials for adjacent superpixels. We have evaluated every method on the same set of features. In line with standard practice in segmentation, we report results with respect to two metrics. The "global" metric counts pixel-wise accuracy, while the "average" metric counts average per-class pixel-wise accuracies. The latter metric is more informative, as the background is typically much more prevalent than foreground. We have employed both variants in the construction of the structured output loss function, $\Delta(y_i, \tilde{y})$, and have trained different models that have optimized each. The regularization parameter has been set using only the training data. We have additionally trained a variant of the probably submodular model that transductively enforces submodularity constraints on the test set as well. As the constraints on the training set guarantee submodularity on the test set with high probability, we do not observe an increase in performance by enforcing the additional constraints. Results are presented in Figure 2. We consider a SVM as the special case of the learning framework considered here where set of constraints is

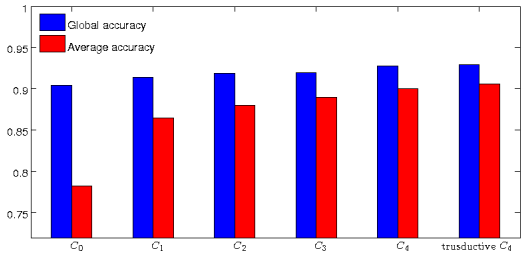
$$\mathcal{C}_0 \equiv \{w \mid w_{00} = w_{11} = w_{12} = w_{21} = \mathbf{0}\}. \qquad (18)$$

We have that $\mathcal{C}_0 \subset \mathcal{C}_1 \subset \mathcal{C}_2 \subseteq \mathcal{C}_3 \subseteq \mathcal{C}_4$. We have verified that the results shown in Figure 2 are statistically significant. A t-test yields a p-value less than 10^{-5} for a comparison of the methods resulting from optimizing the SSVM subject to $w \in \mathcal{C}_1$ vs. $w \in C_4$ both for the "global" and "average" metrics.

Hard constraints that are active at convergence indicate that the data term in the optimization objective is pushing the vector w towards a solution that yields a non-submodular constraint. Table 1 presents the number of active constraints for every method. All methods converge within a similar number of cutting plane iterations: 126 ± 21. Probably submodular models trade off inference error with model expressivity. Table 2 shows the percentage of pairwise constraints that are non-submodular in the test



(a) Accuracies for models trained to optimize the sum of pixel errors over all training images



(b) Accuracies for models trained to optimize the average per-class pixel accuracies.

Figure 2: Comparison of results of the binary segmentation of the TU Darmstadt Database of cows. Models have increased capacity as the plot moves from left to right.

	Non-submodular potentials
$\mathcal{C}_3$	0.5%
$\mathcal{C}_4$	0%

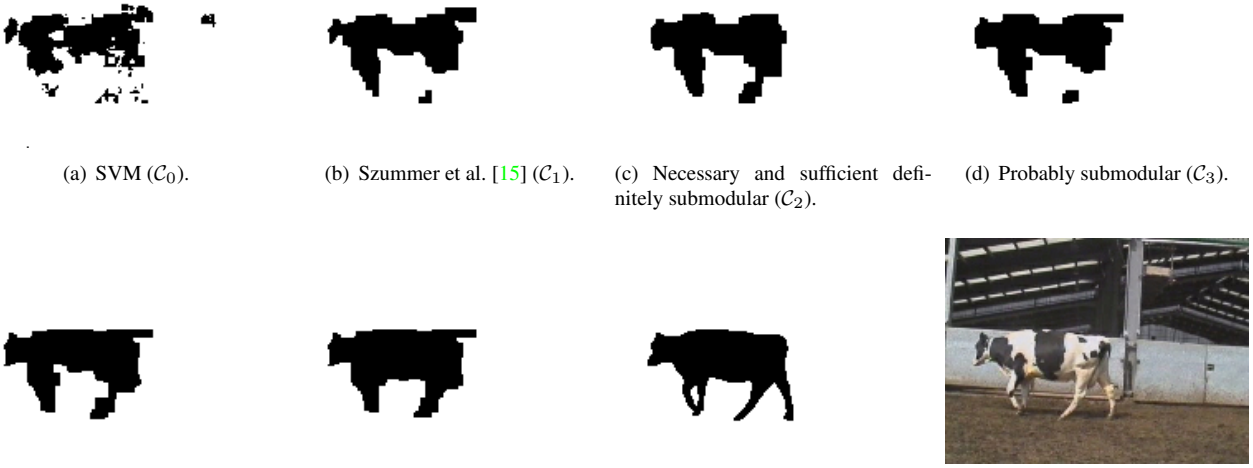
Table 2: The percentage of non-submodular potentials on the test data for probably submodular models. For definitely submodular models this ratio is always equal to zero.

set. Figure 3 gives examples of segmentations predicted by the method of Szummer et al. and by optimization with the probably submodular constraint set $\mathcal{C}_4$.

6. Discussion

The results in Figure 2 indicate that increased model capacity substantially increases accuracies in an image segmentation. Viewed another way, definitely submodular constraints restrict the model capacity to the point that it results in a substantial performance penalty. In contrast, if we rely on probably submodular constraints, we probabilistically guarantee tractability of inference on the test set while enabling more accurate models.

The number of active constraints (Table 1) shows that probably submodular optimization, although globally optimizing over a *much* larger set of linear constraints, in fact



(a) SVM ($\mathcal{C}_0$). (b) Szummer et al. [15] ($\mathcal{C}_1$). (c) Necessary and sufficient definitely submodular ($\mathcal{C}_2$). (d) Probably submodular ($\mathcal{C}_3$).

(e) Probably submodular ($\mathcal{C}_4$). (f) Probably submodular with transductive constraints ($\mathcal{C}_4$). (g) Ground truth. (h) Original image.

Figure 3: Example segmentations from methods trained to optimize each of the constraint sets considered here. As we move from (a) to (e) we increase model capacity and substantially increase the accuracy of the resulting segmentation. .

generates only a small multiple of the constraints required in definitely submodular optimization. We note that this table shows the number of hard constraints for tractability plus the number of one-slack data constraints generated during training of the SSVM.

The empirical evaluation of the number of non-submodular pairwise potentials on the test set (Table 2) validates the theory developed in Section 3. A very small fraction of pairwise potentials were non-submodular, which decreases in the size of the training set. Finally, Figure 3 validates that the improvement in numerical accuracy also results in a qualitative improvement in the semantic segmentation.

There are several interesting approaches for future work. One area of research is in improving the bounds explored in Section 3. We have used a relatively conservative bound on the convex hull, while improvements on bounds on the sample complexity of the convex cone is an active area of learning theory. We have considered submodularity constraints to be hard over the training set, but we may consider the case where we allow them to be violated for a small subset of pairwise potentials through the use of slack variables. A relationship to PAC learning could be an interesting result [17]. In this work, we have derived results for probably submodular constraints, which is appropriate for graph cuts optimization in binary models or α-β swap optimization for multiclass learning, but the principle is equally applicable to the linear metric constraints necessary for α-expansion.

In this work, we have explored the tradeoff between model error and inference error in discriminative training of CRF models. We have developed practical algorithms and theoretical guarantees for probably submodular learning, which substantially improves segmentation accuracy.

Additional information can be found on our project webpage: http://homes.esat.kuleuven.be/~mblaschk/projects/learnConstraints/.

Acknowledgements

This work is partially funded by Internal Funds KU Leuven and FP7-MC-CIG 334380. We acknowledge support from the Research Foundation - Flanders (FWO) through project number G0A2716N.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 4

[2] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of Markov random fields for segmentation of 3D scan data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 169–176, 2005. 2

[3] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *FOCS*, pages 616–623, 1999. 2, 3

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 1

[5] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages 105–112, 2001. 2

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001. 2, 3

[7] C. Buchta. Distribution-independent properties of the convex hull of random points. *Journal of Theoretical Probability*, 3(3):387–393, 1990. 2, 4

[8] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995. 2

[9] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010. 2

[10] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984. 2

[11] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, Oct. 2009. 2

[12] A. R. Klivans, P. M. Long, and A. K. Tang. Baums algorithm learns intersections of halfspaces with respect to log-concave distributions. In *In Proc. of the 13th Intl. Workshop on Randomization and Computation (RANDOM*, pages 588–600, 2009. 2, 3

[13] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 1995. 2, 4

[14] S. Nowozin, P. V. Gehler, and C. H. Lampert. On parameter learning in CRF-based approaches to object class image segmentation. In *ECCV*, volume 6, pages 98–111, 2010. 2

[15] M. Szummer, P. Kohli, and D. Hoiem. Learning crfs using graph cuts. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *European Conference on Computer Vision*, volume 5303 of *Lecture Notes in Computer Science*, pages 582–595. Springer, 2008. 2, 3, 6

[16] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005. 2

[17] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. 2, 6